Biological microarray interpretation

Breitling, Rainer

Published in:
Biochimica et Biophysica Acta %28BBA%29 - Gene Structure and Expression

DOI:
10.1016/j.bbaexp.2006.06.003

Publication date:
2006

Available online at www.sciencedirect.com
ScienceDirect
Biochimica et Biophysica Acta 1759 (2006) 319–327
BIOCHIMICA ET BIOPHYSICA ACTA
BBA
www.elsevier.com/locate/bbaexp

Review

Biological microarray interpretation: The rules of engagement

Rainer Breitling *

Groningen Bioinformatics Centre, University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands

Received 8 May 2006; received in revised form 30 June 2006; accepted 30 June 2006
Available online 13 July 2006

Abstract

Gene expression microarrays are now established as a standard tool in biological and biochemical laboratories. Interpreting the masses of data generated by this technology poses a number of unusual new challenges. Over the past few years a consensus has begun to emerge concerning the most important pitfalls and the proper ways to avoid them. This review provides an overview of these ideas, beginning with relevant aspects of experimental design and normalization, but focusing in particular on the various tools and concepts that help to interpret microarray results. These new approaches make it much easier to extract biologically relevant and reliable hypotheses in an objective and reasonably unbiased fashion.

Keywords: Microarray; Gene expression; Bioinformatics; Tool; Computational biology; System biology

1. Introduction

Microarrays are gaining popularity in biological laboratories by the day. In their standard application, they measure the gene expression status of a particular sample, by quantifying the mRNA levels of all genes in highly parallel fashion. A single array hybridization generates as many data as several classical Ph.D. theses taken together. This makes the technology very appealing as a shortcut towards data production, despite the large financial and technical demands. To turn these data into biological insight it is, however, necessary to interpret the collected data. After having come to grips with the manipulation of the large amount of collected data, the interpretation may seem almost trivial at first: there are many possible stories that a creative expert biologist can detect in the data, treating the results for each gene more or less like familiar Northern blot information. But in contrast to more traditional approaches the main challenge is not to come up with an explanation for the behavior of a single gene, but with one that is consistent and well supported in the context of complementary information on thousands of other genes.

New ideas on how to achieve this task are developed rapidly, new microarray tools and specialized microarray statistics are abounding. Making an informed choice between them may sometimes be daunting, but fortunately over the last few years a number of comparative evaluation studies and the accumulated experience from many thousands of microarray studies have lead to the emergence of a set of guidelines and informal standards that help in the process. In the following sections we will first outline the basic conditions that need to be fulfilled before a successful interpretation can start and describe how to detect differentially expressed genes and how to avoid the major pitfall associated with multiple testing. Then we proceed with a highly selective outline of methods to organize the differential expression information and to explore and annotate the results to obtain a biologically coherent interpretation. We conclude with an overview of higher-level tools that support the integration and extension of the array results with additional data sources. Fig. 1 summarizes the interpretation procedure.

2. Biological question and experimental design

A number of critical steps need to be taken before one can think of interpreting a microarray result successfully. One of these is the proper randomization during the experiment. It is

* Tel.: +31 50 363 8088; fax: +31 50 363 7976.
E-mail address: r.breitling@rug.nl.

0167-4781/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbaexp.2006.06.003

often jokingly said that the results of a laboratory experiment depend on the phase of the moon. For microarray analysis this is almost true. Gene expression responds very sensitively to changes in environmental conditions, even if they appear very minor. Consistent and statistically significant fluctuations in expression pattern have been reported in wild type yeast cultures incubated under constant standard growth conditions [1]. Even for genetically homogeneous inbred mouse strains reared under highly-controlled, pathogen-free laboratory conditions and matched for age and sex reproducible and statistically significant inter-individual variation in gene expression has been reported [2]. It was necessary to control many additional variables like social status, stress, and food intake to reduce this biological variation to a minimum. Even then, individual mice showed significant differential expression of some genes. Similar effects have been reported for *Arabidopsis*, where even simply touching the plants can lead to significant changes in gene expression [3]. Therefore, it is of utmost importance for the generation of interpretable microarray results to randomize or control all possible confounding factors. If two conditions are to be compared, then it is not sufficient to, for example, obtain the material at the same temperature and in the same medium, but it should preferably be grown in the same incubator, in random spatial arrangement and harvested in random order by the same person. The exact factors to be randomized will depend on the particular experimental set-up, but in any case it will be crucial to realize that the most unexpected subtle systematic difference between the sampling (and measurement) conditions will lead to biased results—making the data all but impossible to interpret.

To be able to interpret microarray data, it is also necessary to have a sufficient number of replicate measurements. Such replication is necessary to assess which results will have real predictive value, i.e. are expected to be verified in a new experiment, and which observations are only spurious. Statistical approaches allow one to assess the number of replicates necessary to reliably detect an expected effect of a certain size. For most applications, where financial and logistic constraints limit the number of hybridizations, this is merely an academic exercise. The more replicates the better. A real choice occurs when the type of replicates is chosen. Here it is important to realize that arrays show surprisingly little technical (labeling- or hybridization-specific) variation. Most of the observed expression variation is due to biological variability (day-to-day or interindividual variation; [4–7]). It is, therefore, most efficient and informative to perform as many *biological* replicates as possible. Rather than hybridizing a specific sample twice, isolate the same type of mRNA from a completely independent new sample. This will help to detect those expression changes that are reproducible and relevant. Detailed reviews of experimental design considerations, in particular for more complex experiments such as large time courses or multifactorial comparisons, are available in [8,9].

3. Hybridization

The choice of array platform and other technical details of the actual microarray experiment (sample preparation, hybridization, image processing) of course also influence the downstream analysis, but these will not be discussed in this paper. For

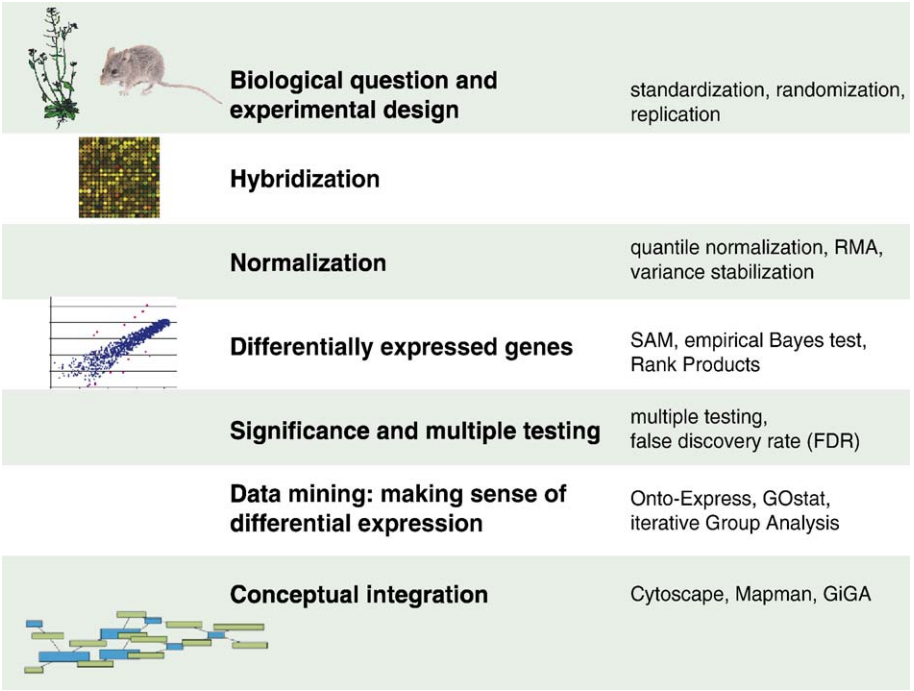

Fig. 1. Critical steps in microarray interpretation. The major steps of the usual interpretation process are outlined. Major pitfalls and some related analytical approaches are briefly indicated at the right. See main text for detailed discussion and the description of additional relevant tools and methods.

recent reviews and comparisons of various technologies see [6,10–13].

4. Normalization

Before analysis, data usually need to be normalized, to remove obvious systematic differences between samples. For example, in a standard array experiment, the amount of labeled cRNA in all samples should be the same, so normalization could involve some scaling of the expression values, so that the average signal is identical for all arrays. Many procedures have been suggested for systematic normalization, but it seems that the simplest strategies often perform best. For two-color arrays, normalization can, for example, be performed by a simple quantile normalization [14], which makes the distribution of intensities identical on all arrays and colors. This extends the original (and still widely used) idea that one could normalize by making the mean and variance identical for all samples and has been reported to have superior performance [14]. This can then be followed by a variance stabilizing normalization [15–17], which handles the problem caused by wide fluctuations of log-ratios for genes that are expressed at the detection limit. A simple alternative – that often works surprisingly well – is just to forgo background subtraction (which in any case just tends to introduce additional noise). This procedure has a similar effect as a started-log transformation, which is the most simple form of variance stabilization [18]. For one-color genechip arrays, it seems most important *not* to use the standard normalization provided by the MAS5/GCOS software of Affymetrix. This normalization has been shown to perform consistently poorly in various comparative evaluations [19–23], so use of MAS5/GCOS should be restricted to the lowest level analysis, such as the initial image processing. Of the alternative normalization techniques, RMA [19] seems to show the most consistent good performance, with its modification GCRMA [24] and the dChip method [25] being close second. The RMA method includes quantile normalization and variance stabilization components and doesn't use the "information" from mismatch probes, so it corresponds quite closely to the procedure suggested for two-color arrays. Both methods do not do background subtraction and thus do not have to deal with negative expression values. Avoiding arbitrary filtering, e.g. based on above-background expression or spot quality, is usually recommended.

5. Differentially expressed genes

Detecting genes that show reproducible differences in mRNA abundance between sample classes is probably the most basic and important step of microarray analysis. A huge variety of methods have been used for this task, and when sufficient replicates are available all of them work. What is not sufficient is a simple calculation of fold-change and arbitrary thresholding—unless the data are not replicated at all (which is in itself very problematic). This will fail to discriminate between changes that are spurious and those that are reproducible. Also the choice of a fold-change threshold is very subjective and is so much dependent on the experiment and technological variables that any (or no) choice can be justified. When the number of replicates is very small, methods based on the *t*-test, including the popular SAM [26], have difficulties and can be replaced by equally simple approaches that are designed for small numbers of replicates, such as Rank Products [27,28]. Rank Products have the additional benefit that they by design are able to detect rare but interesting expression changes (with statistical significance), that would be missed by many other approaches. Alternatively, the Empirical Bayes modification of the *t*-test can be used with good success [29,30], as can be other Bayesian approaches [31,32].

For experiments that compare more than two conditions the *t*-test can be replaced by an Analysis of Variance (ANOVA) [33]. This requires some more statistical expertise, but the necessary tools are available with extensive documentation, e.g., in the BioConductor package of the R statistical language [34].

In any case it is highly recommended to use "standard" methods, instead of reinventing the wheel. It is very challenging to demonstrate that a new method really outperforms the available ones. This needs to be shown quantitatively for a large number of datasets. Any performance differences will be subtle and difficult to show robustly, and the danger of "optimizing" (over-fitting) an analysis technique for a particular dataset is large.

The result of all the statistical tests is a *p*-value for each gene, which basically describes how likely it would be to observe a particular differential expression by chance. This concept is not just a formal statistical idea, but also a straightforward estimate of the ability to predict the outcome of future experiments. This estimate is based on the observed variability and effect size in replicated independent measurements. Therefore replicate measurements are essential in any microarray experiment.

6. Significance and multiple testing

Multiple testing is the most dangerous pitfall of the microarray interpretation process. It comes in many forms, and is sometimes so cleverly disguised that it is hard to recognize the problem. In the classical case, it refers to the situation that one tests the differential expression of thousands of genes in parallel. The classically calculated statistical significance (*p*-values) are designed for single tests, they tell the user how likely a certain effect is to be observed by chance. For example, a *p*-value of 0.05 means that such a strong effect (e.g., differential expression) would be observed by chance in one of 20 tests. Usually that's considered significant, but in the multiple testing situation of microarrays, it rarely is. Among 10,000 genes tested in parallel, it is expected that about 500 genes would reach such a *p*-value just by chance, even if they are not differentially expressed at all. These genes would be likely to be false positives, basically artifacts of the statistical analysis procedure.

One way to overcome this problem is to use stricter thresholds. In the above example, a *p*-value smaller than $0.05/10{,}000 = 5 * 10\text{E-}6$ would be expected only once in 20 experiments, just as strict as the original single-test

threshold. This so-called Bonferroni correction [35] is, however, quite strict and is currently rarely applied in microarray interpretation. In its place, the False Discovery Rate (FDR; [36]) is most widely used. This clever and intuitive method controls the expected number of false-positive results in the list of results, rather than the number of experiments in which any mistake is made. This leads to a more lenient threshold that also flexibly depends on the observed data. The FDR is so simple that one can estimate it manually for a given result, and that can be highly instructive when reading the literature. For example, if a paper reports that 25,000 genes were analyzed, 500 of which were differentially expressed at a p-value of 0.01, one can easily calculate that $N = 25{,}000 \times 0.01 = 250$ genes are expected to show such a change by chance, leading to an FDR of 50%, much larger than would ever be acceptable in a conventional publication. For microarray publications, FDRs between 1% and 10% are considered reasonably strict. More lenient thresholds lead to results that are difficult to distinguish from random gene lists. Notice that there are other variants of multiple testing, not just the large number of genes causes problems, but sometimes tests are also performed multiple times for various subdivisions of the data, for particular versions of the normalization or for different test statistics. Each of these has the potential to lead to the discovery of spurious false positives if it is not treated carefully.

Any statistical significance test has the potential to reject true positive results, i.e. it will miss genes that are in fact differentially expressed. To control this problem, it is useful to complement the FDR estimates with information on the expected false negative rate [37] or 'miss rate' [38].

The FDR is a property of an entire list of genes, it gives an estimate of how many of the genes that are reported as differentially expressed are likely to be false positives. Of course within this list of genes, those with the smaller p-values will still be the most significant candidates, so the classical p-values are still relevant (or one may want to use their FDR-related equivalent, the gene-specific q-values [39]). Also, among the genes that are statistically significant at a certain FDR level, some may be unlikely candidates, for instance because their overall fold-changes are very low. Additional limitation of the reported list based on such criteria may be helpful for display purposes.

As long as multiple testing is properly addressed, all choices at the other steps of the interpretation process have only relatively minor consequences. Ignoring the multiple testing problem has, on the other hand, resulted in the publication of elaborate biological stories that were based on about 100% false positive expression changes. That it is possible to generate such stories is a testament to the creativity and expert knowledge of the analyst—but it should also serve as a warning for anybody trying to interpret a microarray result. It is very easy to see the patterns one wants to see and assign meaning to observations that are purely the result of random fluctuations. Because microarrays generate so many data, they make it particularly easy to come up with seemingly plausible, but irreproducible, interpretations [40]. The next section discusses some approaches that try to minimize these dangers.

7. Data mining: making sense of differential expression

To interpret microarrays in a systematic way, some "dimensionality reduction" is required. This means that the factors or phenomena that dominate the observed data need to be extracted and emphasized. There are two principal ways of achieving this, both with their advantages and disadvantages. The first one, cluster analysis, is exclusively data-driven, the second one, annotation analysis, is knowledge-driven. The first one may tend to miss subtle patterns that are immediately obvious if previous biological knowledge is considered, the second one may be overly restrictive (or even biased) by its almost exclusive focus on genes and processes that have been annotated already. We will discuss both approaches in turn.

Clustering and its relatives (like PCA) are almost emblematic for microarray analysis. A gene expression paper is almost considered incomplete if it doesn't contain a colored heatmap sorted by hierarchical clustering as first introduced in a classical paper by Eisen et al. [41]. These techniques have been traditionally used for dimensionality reduction, i.e. to make the huge amounts of complex information contained in a microarray more digestible for the interpreter. They are extremely useful for visualization purposes and quality control. For example, a heatmap sorted by hierarchical clustering will give a quick visual impression of which groups of samples are most similar and highlight samples that stand out from their group (and thus may have technical problems). It will also show if there are large groups of genes that show consistent correlated behavior across samples. However, because they focus on the large-scale patterns distinguishing samples, a limited number of dimensions, they may discard too much of the more subtle variation that may contain most of the biological signal. On the other hand, in some situations they may focus too much on stochastic fluctuations in the data. If one is comparing two conditions in replicated measurements, clustering cannot produce more information than is already available from the differential expression analysis. Any clusters beyond the simple up-down and down-up patterns will be based on the random fluctuations within conditions. Where the data are suitable for clustering approaches, i.e. where a sufficiently large number of conditions are compared, it may be interesting to consider more flexible approaches than the most popular hierarchical clustering and PCA. For example, k-means clustering may perform consistently better than hierarchical clustering [42,43] and self-organizing maps can also be very powerful if applied with sufficient statistical expertise [44]. Other approaches, like Linear Factor Models (LFM, [45]) allow each gene to be a member of multiple clusters and due to its biology-based underlying assumptions generates results that are much more accurate reflections of the physiological pattern than is possible with PCA. Using a biologically plausible model the LFM approach will basically detect the most suitable small number of "biological processes" (e.g. activated regulatory cascades) that are necessary to give an accurate representation of the observed

expression patterns. The signature algorithm [46] is based on similar concepts and is particularly suited for the (meta-) analysis of very large expression datasets. If the dataset is comprehensive enough it can also be used to determine transcriptional regulatory modules, as has been shown for a 1000-array dataset from yeast [47]. But, while it is useful to consider clustering approaches that go beyond the traditional, simplicity of the analysis can also be beneficial: for the related task of classification Dudoit et al. [48] have demonstrated that the simplest methods show remarkably good performance.

If clusters are sufficiently coherent (i.e. specifically regulated by a common set of transcription factors), they can be used to mine for the responsible transcription factor modules in their 3′ untranslated regions [49]. One tool that is useful for identifying overrepresented motifs is MEME [50]. The false positive rate of such methods is very high and care has to be taken to validate the discovered motifs. Even if no additional experimentation is immediately possible, examining the results for low complexity regions, which may correspond to compositional biases rather than transcription factor binding sites, is an important first step.

The next step, with or without clustering, is the annotation of the results. This can be done manually, based on expert knowledge and careful literature study on the dozens or hundreds of genes that are detected. This step is most imperiled by interpreter bias, a.k.a. expert knowledge and previous expectations. To overcome this problem, automated methods have been devised to perform this task. There is a very large number of tools available [51] that differ in the details of their implementation (e.g., which gene identifiers can be processed, how the results are displayed, how the annotation is kept up-to-date), but they all use the same basic principle. Given a list of genes and the annotation of the genome they calculate if a certain functional category is statistically overrepresented in the selected group of genes, compared to the rest of the genome. The most surprisingly enriched gene groups usually yield valuable clues with respect to the affected biological processes in a particular experiment. Because the enrichment is calculated on statistical principles and summarized in a regular *p*-value, the results are more objective and easier to compare than those of a manual analysis (which may often lead to similar conclusions). The most popular software tools used for this purpose are Onto-Express [52], GOstat [53], GoMiner [54] and EASE [55], but many others are available and may be just as useful or even better for particular situations (see review by Khatri and Draghici, [51] for a detailed comparison of a large part of them).

The annotation that is most often used for this type of analysis is the Gene Ontology [56], a controlled vocabulary that uses standardized and centrally curated terms to describe the biological function of genes and their products. Such a well-defined vocabulary is on the one hand useful, because it makes results more comparable, but on the other hand it can be overly restrictive and can miss important information that would be provided by more flexible annotation systems. This is a major limitation of many of the tools discussed above. Using custom-made annotations, such as groups of genes that one knows to be of interest, based on literature research or previous experiments of various kinds, can be a major advantage. It does, however, often require at least a moderate amount of bioinformatics support.

Another limitation of the gene-enrichment detection methods discussed, is their focus on a pre-defined, fixed group of differentially expressed genes. Of course, even if the selection of this group is based on solid statistical criteria, it is influenced by arbitrary decisions: which *p*-value threshold should be used? How many false positive results is one willing to accept? But also, and this is sometimes neglected, how many false negative results follow from this decision? One implementation of the gene-enrichment detection algorithm, called iterative Group Analysis [57], overcomes this problem by using flexible thresholds, so that for each functional group of genes, the most appropriate cutoff point is defined automatically. This can lead to dramatic improvements for the detection of some functional classes, in particular those that show concerted and very interesting changes in expression, but only limited absolute changes. iGA can also handle a wide range of custom-defined gene annotations.

A global analysis of gene lists that determines the overall enrichment of genes is only a first step in the annotation process. The next level of analysis requires to make connections between genes. Are the detected enriched processes related? How do genes and their functions link up to a larger picture? Again this placement of the isolated findings in their physiological context can be done manually, with full creative freedom, but it can also be supported and guided by a number of computational tools that help to visualize the relationships between genes based on complementary knowledge.

A successful tool of this type is the popular MAPMAN software [58,59], which is mainly used in the plant science community. It allows the projection of gene expression data on network pictures, such as metabolic maps and signaling pathways, and can serve to rapidly highlight areas in cellular physiology that are most strongly affected by differential expression. It also can be used to create aesthetically pleasing, publication-quality figures of these network pictures.

Cytoscape [60] is another program for integrating gene interaction information and gene expression data in a unified framework for visualization and exploration. It allows the display of large gene networks, which place the individual expression changes in their biological context. The program has been extended by a wide variety of specialized modules, both by the original authors and other groups. An example is the ActiveModules plug-in, which can be used to detect connected subnetworks within a gene network whose members show significant coordinated changes in mRNA-expression over a variety of experimental perturbations [61]. This extensible architecture of the software and its excellent graph visualization properties make it very versatile for integrating data, including many types of molecular interaction data, and working in a systems biology framework.

Graph-based iterative group analysis (GiGA; [62]) is another higher-level interpretation method. It extends the scope of iGA (see above) to detect statistically significant areas (subgraphs) in a gene network, which respond most remarkably to a certain

condition. The connections in the network can be based on gene regulation data (including transcription factor binding sites), metabolic connectivity, physical interactions, literature co-citations or other information that can serve to link genes conceptually [63]. Focusing on those areas in the conceptual network that are most strongly affected by the treatment of interest can narrow down the search space sufficiently to arrive at a consistent and convincing interpretation rapidly.

A related tool is Pathway-Express, which is restricted to a single list of potential genes of interest [64]. Its aim is to identify and display those cellular pathways involving genes of the list that are most "interesting" based on probabilistic criteria. GeneMAPP and the related MAPPFinder gene ontology tool also fall into this category of interpretation tools [65,66].

The tools described will help to find biological processes that seem to be responding to certain treatments or conditions. The results must, however, be taken with a grain of salt. It is not uncommon to detect, e.g., ribosomal proteins as a responsive group. Does that mean that protein synthesis is noticeably changed? Possibly, although protein and mRNA levels correlate only weakly for ribosomal proteins [67]. But will that be part of a specific adaptive program? That is even less likely. It seems that certain groups of genes, and ribosomal ones in particular, respond very sensitively to any fluctuation in the environmental conditions, but probably without major specific consequences. The fact that ribosomal expression changes have been reported for dozens of experiments, ranging from yeast diauxic shift [68] to metamorphosis of ants [69], makes such observations too unspecific for further interpretation, even if they have a general biological relevance (energy expenditure for protein synthesis should, after all, be finely regulated). This broader context of expression variation needs to be kept in mind when analyzing a particular experiment. It also may be necessary to consider the possibility that most expression variations may be random and of no biological significance, no matter how statistically significant they may be. They may be due to spontaneous mutations of regulatory sites that become fixed in the genome by chance, but may be inconsequential and entirely unrelated to the adaptive response to a particular physiological situation. On the other hand, it has been shown for *C. elegans* that intraspecific evolution of transcriptional variation is in many cases subject to intense stabilizing selection [70]. This observation might indicate that a considerable fraction of expression changes will have adaptive consequences and is therefore biologically relevant.

Comparing gene expression responses in various conditions, e.g. the response of mutant and wild type cells to a common stress drug treatment, is also often of interest. Here one can't directly compare the gene expression levels (wild type and mutant will differ even without treatment), but will instead focus on the relative differences in response compared to the corresponding unchallenged state. The statistically comprehensive way of analysis such a case is analysis of variance (ANOVA; [33]), but this requires statistical expertise and the results are not always straightforward to visualize and interpret. The more common method is to plot Venn diagrams that show the overlap and difference between the sets of differentially expressed genes. The disadvantage is that the message (and interpretation) of Venn diagrams can be dramatically different if different thresholds are used for defining the set of differentially expressed genes. VectorAnalysis [71] is a tool that offers the intuitive visualization and grouping of genes into various response classes, as in Venn diagrams, but at the same time provides a statistical foundation of these assignments and is independent of arbitrary expression thresholds.

8. Conceptual integration

Several tools help to move beyond the interpretation of microarray results in isolation. Two particularly comprehensive examples are the commercial bioinformatics efforts of Genomatix and Ingenuity. The former integrates expression data with up-to-date genome information, offers powerful tools for extracting potential regulatory motifs (transcription factor modules; [72]) and also gives targeted access to the literature associated with a particular set of genes, using statistical methods to detect potential links between publications and genes [73]. The hierarchical combination of the components of this toolbox can highlight relevant cross-connections that may escape a more compartmentalized analysis. At the same time they offer so many potential leads that wise restraint is necessary before making hard conclusions about the meaning and significance of a particular observation—the statistics implemented in the platform are only a first step to help with this, but careful consideration of hidden multiple testing "traps" is required. How many similarly "significant" results were discarded, because they didn't add up to an exciting biological story? The Ingenuity pathway analyzer uses similar basic principles as the GiGA approach described above [62], but complements it with extensively curated gene networks and tools to add and edit customized networks (www.ingenuity.com). The user-interface also provides easy access to the available literature, both relating to genes and to the connections between them. Both software packages are web-based for easy access and offer free trial accounts and special licensing agreements for academic scientists. Separate free software for many of the individual steps of the analysis offered by these commercial efforts is available, but they derive their main power from the integration of diverse information in a unified and highly accessible interface.

Of course, data integration and the generation of complex, informative hypotheses at the systems level is not an aim in itself. Only if the hypotheses are sufficiently convincing and reliable to be tested in the biological laboratory has the interpretation process been really successful. In fact, restricting the reported results of a microarray study to those aspects that one is willing to pursue by direct experimental testing, might yield an even greater improvement in interpretation quality than refinements of subtle statistical aspects.

Another important aspect of microarray interpretation is its open nature. The interpretation is not finished when the original experimenter has made her conclusions, but the data can be reanalyzed by other people. Comparing one's own results to those of other people can lead to important refinements and

confirmations of a particular array interpretation. For example, a *C. elegans* researcher may want to find the place of their most responsive genes on the global gene expression map of Kim et al. [74], which it possible via the website of the authors. To make such comparative analysis possible, it is essential that the data are publicly available. The current major repositories (Gene Expression Omnibus [75], ArrayExpress [76], Stanford Microarray Database [77]) are still very difficult to exploit spontaneously and in a global manner, but the user interfaces are continuously developing and already give access to tens of thousands of experiments. Every researcher should consider to make their own data available in this form for future interpretation.

9. Conclusion

Microarray interpretation is as much an art as it is a science. There are, however, a number of standards that have emerged over the past years that help to make the interpretation process more objective and reliable. The most important advance is the availability of tools that reduce the human factor in the initial annotation of the experimental results. This does not mean that expert analysts and interpreters will be replaced by the computer, but opens new opportunities to produce creative yet well-founded interpretations to hold up against the mounting onslaught of new and exciting microarray datasets.

Acknowledgements

I am grateful to Pawel Herzyk, Patrick Armengaud and three anonymous referees for their constructive comments on the manuscript.

References

[1] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, et al., Functional discovery via a compendium of expression profiles, Cell 102 (1) (2000) 109–126.

[2] M. Seltmann, M. Horsch, A. Drobyshev, Y. Chen, M.H. de Angelis, J. Beckers, Assessment of a systematic expression profiling approach in ENU-induced mouse mutant lines, Mamm. Genome 16 (1) (2005) 1–10.

[3] T. Chotikacharoensuk, R.N. Arteca, J.M. Arteca, Use of differential display for the identification of touch-induced genes from an ethylene-insensitive Arabidopsis mutant and partial characterization of these genes. J. Plant Physiol. (in press).

[4] J.J. Chen, R.R. Delongchamp, C.A. Tsai, H.M. Hsueh, F. Sistare, K.L. Thompson, V.G. Desai, J.C. Fuscoe, Analysis of variance components in gene expression data, Bioinformatics 20 (9) (2004) 1436–1446.

[5] S.A. van Hijum, A. de Jong, R.J. Baerends, H.A. Karsens, N.E. Kramer, R. Larsen, C.D. den Hengst, C.J. Albers, J. Kok, O.P. Kuipers, A generally applicable validation scheme for the assessment of factors involved in reproducibility and quality of DNA-microarray data, BMC Genomics 6 (1) (2005) 77.

[6] C.L. Yauk, M.L. Berndt, A. Williams, G.R. Douglas, Comprehensive comparison of six microarray technologies, Nucleic Acids Res. 32 (15) (2004) e124.

[7] S.O. Zakharkin, K. Kim, T. Mehta, L. Chen, S. Barnes, K.E. Scheirer, R.S. Parrish, D.B. Allison, G.P. Page, Sources of variation in Affymetrix microarray experiments, BMC Bioinformatics 6 (2005) 214.

[8] M.K. Kerr, G.A. Churchill, Statistical design and the analysis of gene expression microarray data, Genet. Res. 77 (2) (2001) 123–128.

[9] J.P. Townsend, J.W. Taylor, Designing experiments using spotted microarrays to detect gene regulation differences within and among species, Methods Enzymol. 395 (2005) 597–617.

[10] S. Draghici, P. Khatri, A.C. Eklund, Z. Szallasi, Reliability and reproducibility issues in DNA microarray measurements, Trends Genet. 22 (2) (2006) 101–109.

[11] G. Hardiman, Microarray platforms—Comparisons and contrasts, Pharmacogenomics 5 (5) (2004) 487–502.

[12] R.A. Irizarry, D. Warren, F. Spencer, I.F. Kim, S. Biswal, B.C. Frank, E. Gabrielson, J.G. Garcia, J. Geoghegan, G. Germino, et al., Multiple-laboratory comparison of microarray platforms, Nat. Methods 2 (5) (2005) 345–350.

[13] J.E. Larkin, B.C. Frank, H. Gavras, R. Sultana, J. Quackenbush, Independence and reproducibility across microarray platforms, Nat. Methods 2 (5) (2005) 337–344.

[14] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2) (2003) 185–193.

[15] B.P. Durbin, J.S. Hardin, D.M. Hawkins, D.M. Rocke, A variance-stabilizing transformation for gene-expression microarray data, Bioinformatics 18 (Suppl. 1) (2002) S105–S110.

[16] B.P. Durbin, D.M. Rocke, Variance-stabilizing transformations for two-color microarrays, Bioinformatics 20 (5) (2004) 660–667.

[17] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, M. Vingron, Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics 18 (Suppl. 1) (2002) S96–S104.

[18] D.M. Rocke, B. Durbin, Approximate variance-stabilizing transformations for gene-expression microarray data, Bioinformatics 19 (8) (2003) 966–972.

[19] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 4 (2) (2003) 249–264.

[20] R.A. Irizarry, Z. Wu, H.A. Jaffee, Comparison of Affymetrix GeneChip expression measures, Bioinformatics 22 (7) (2006) 789–794.

[21] L.X. Qin, R.P. Beyer, F.N. Hudson, N.J. Linford, D.E. Morris, K.F. Kerr, Evaluation of methods for oligonucleotide array data via quantitative real-time PCR, BMC Bioinformatics 7 (2006) 23.

[22] K. Shedden, W. Chen, R. Kuick, D. Ghosh, J. Macdonald, K.R. Cho, T.J. Giordano, S.B. Gruber, E.R. Fearon, J.M. Taylor, et al., Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data, BMC Bioinformatics 6 (1) (2005) 26.

[23] L.M. Cope, R.A. Irizarry, H.A. Jaffee, Z. Wu, T.P. Speed, A benchmark for Affymetrix GeneChip expression measures, Bioinformatics 20 (3) (2004) 323–331.

[24] Z. Wu, R. Irizarry, R. Gentleman, F. Murillo, F.A. Spencer, A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Technical Report, John Hopkins University, Department of Biostatistics Working Papers, Baltimore, MD 2003.

[25] Z. Wu, M.S. Siadaty, G. Riddick, H.F. Frierson Jr., J.K. Lee, W. Golden, S. Knuutila, G.M. Hampton, W. El-Rifai, D. Theodorescu, A novel method for gene expression mapping of metastatic competence in human bladder cancer, Neoplasia 8 (3) (2006) 181–189.

[26] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U. S. A. 98 (9) (2001) 5116–5121.

[27] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, FEBS Lett. 573 (1–3) (2004) 83–92.

[28] R. Breitling, P. Herzyk, Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data, J. Bioinform. Comput. Biol. 3 (5) (2005) 1171–1189.
[29] P. Baldi, A.D. Long, A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes, Bioinformatics 17 (6) (2001) 509–519.
[30] B. Efron, R. Tibshirani, J.D. Storey, V. Tusher, Empirical Bayes analysis of a microarray experiment, J. Am. Stat. Assoc. 456 (2001) 1151–1160.
[31] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (1) (2004) (Article3).
[32] J.P. Townsend, D.L. Hartl, Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments, Genome Biol. 3 (12) (2002) (RESEARCH0071).
[33] M.K. Kerr, M. Martin, G.A. Churchill, Analysis of variance for gene expression microarray data, J. Comput. Biol. 7 (6) (2001) 819–837.
[34] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al., Bioconductor: open software development for computational biology and bioinformatics, Genome Biol. 5 (10) (2004) R80.
[35] C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8 (1936) 3–62.
[36] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. 57 (1) (1995) 289–300.
[37] A.W. Norris, C.R. Kahn, Analysis of gene expression in pathophysiological states: balancing false discovery and false negative rates, Proc. Natl. Acad. Sci. U. S. A. 103 (3) (2006) 649–653.
[38] J. Taylor, R. Tibshirani, B. Efron, The 'miss rate' for the analysis of gene expression data, Biostatistics 6 (1) (2005) 111–117.
[39] J.D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proc. Natl. Acad. Sci. USA 100 (16) (2003) 9440–9445.
[40] J.P. Ioannidis, Microarrays and molecular research: noise discovery? Lancet 365 (9458) (2005) 454–455.
[41] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. U. S. A. 95 (25) (1998) 14863–14868.
[42] S. Datta, S. Datta, Comparisons and validation of statistical clustering techniques for microarray gene expression data, Bioinformatics 19 (4) (2003) 459–466.
[43] K.Y. Yeung, D.R. Haynor, W.L. Ruzzo, Validating clustering for gene expression data, Bioinformatics 17 (4) (2001) 309–318.
[44] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proc. Natl. Acad. Sci. U. S. A. 96 (6) (1999) 2907–2912.
[45] M. Girolami, R. Breitling, Biologically valid linear factor models of gene expression, Bioinformatics 20 (17) (2004) 3021–3033.
[46] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, N. Barkai, Revealing modular organization in the yeast transcriptional network, Nat. Genet. 31 (4) (2002) 370–377.
[47] J. Ihmels, S. Bergmann, N. Barkai, Defining transcription modules using large-scale gene expression data, Bioinformatics 20 (13) (2004) 1993–2003.
[48] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc. 97 (457) (2002) 77–87.
[49] T. Werner, Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data, Pharmacogenomics 2 (1) (2001) 25–36.
[50] T.L. Bailey, C. Elkan, The value of prior knowledge in discovering motifs with MEME, Proc. Int. Conf. Intell. Syst. Mol. Biol. 3 (1995) 21–29.
[51] P. Khatri, S. Draghici, Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinformatics 21 (18) (2005) 3587–3595.
[52] P. Khatri, S. Draghici, G.C. Ostermeier, S.A. Krawetz, Profiling gene expression using onto-express, Genomics 79 (2) (2002) 266–270.
[53] T. Beissbarth, T.P. Speed, GOstat: find statistically overrepresented Gene Ontologies within a group of genes, Bioinformatics 20 (9) (2004) 1464–1465.
[54] B.R. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, et al., GoMiner: a resource for biological interpretation of genomic and proteomic data, Genome Biol. 4 (4) (2003) R28.
[55] D.A. Hosack, G. Dennis Jr., B.T. Sherman, H.C. Lane, R.A. Lempicki, Identifying biological themes within lists of genes with EASE, Genome Biol. 4 (10) (2003) R70.
[56] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (1) (2000) 25–29.
[57] R. Breitling, A. Amtmann, P. Herzyk, Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments, BMC Bioinformatics 5 (2004) 34.
[58] O. Thimm, O. Blasing, Y. Gibon, A. Nagel, S. Meyer, P. Kruger, J. Selbig, L.A. Muller, S.Y. Rhee, M. Stitt, MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes, Plant. J. 37 (6) (2004) 914–939.
[59] B. Usadel, A. Nagel, O. Thimm, H. Redestig, O.E. Blaesing, N. Palacios-Rojas, J. Selbig, J. Hannemann, M.C. Piques, D. Steinhauser, et al., Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses, Plant. Physiol. 138 (3) (2005) 1195–1204.
[60] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.
[61] T. Ideker, O. Ozier, B. Schwikowski, A.F. Siegel, Discovering regulatory and signalling circuits in molecular interaction networks, Bioinformatics 18 (Suppl. 1) (2002) S233–S240.
[62] R. Breitling, A. Amtmann, P. Herzyk, Graph-based iterative Group Analysis enhances microarray interpretation, BMC Bioinformatics 5 (2004) 100.
[63] M. Benson, R. Breitling, Network theory to understand microarray studies of complex diseases. Current Molecular Medicine (in press).
[64] P. Khatri, S. Sellamuthu, P. Malhotra, K. Amin, A. Done, S. Draghici, Recent additions and improvements to the Onto-Tools, Nucleic Acids Res. 33 (2005) W762–W765 (Web Server issue).
[65] K.D. Dahlquist, N. Salomonis, K. Vranizan, S.C. Lawlor, B.R. Conklin, GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, Nat. Genet. 31 (1) (2002) 19–20.
[66] S.W. Doniger, N. Salomonis, K.D. Dahlquist, K. Vranizan, S.C. Lawlor, B. R. Conklin, MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, Genome Biol. 4 (1) (2003) R7.
[67] T.J. Griffin, S.P. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood, R. Aebersold, Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*, Mol. Cell Proteomics 1 (4) (2002) 323–333.
[68] J.L. De Risi, V.R. Iyer, P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (5338) (1997) 680–686.
[69] M.A. Goodisman, J. Isoe, D.E. Wheeler, M.A. Wells, Evolution of insect metamorphosis: a microarray-based study of larval and adult gene expression in the ant *Camponotus festinatus*, Evolution Int. J. Org. Evolution 59 (4) (2005) 858–870.
[70] D.R. Denver, K. Morris, J.T. Streelman, S.K. Kim, M. Lynch, W.K. Thomas, The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*, Nat. Genet. 37 (5) (2005) 544–548.
[71] R. Breitling, P. Armengaud, A. Amtmann, Vector analysis as a fast and easy method to compare gene expression responses between different experimental backgrounds, BMC Bioinformatics 6 (2005) 181.

[72] A. Klingenhoff, K. Frech, K. Quandt, T. Werner, Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity, Bioinformatics 15 (3) (1999) 180–186.
[73] M. Seifert, M. Scherf, A. Epple, T. Werner, Multievidence microarray mining, Trends Genet. 21 (10) (2005) 553–558.
[74] S.K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, G.S. Davidson, A gene expression map for *Caenorhabditis elegans*, Science 293 (5537) (2001) 2087–2092.
[75] T. Barrett, T.O. Suzek, D.B. Troup, S.E. Wilhite, W.C. Ngau, P. Ledoux, D. Rudnev, A.E. Lash, W. Fujibuchi, R. Edgar, NCBI GEO: mining millions of expression profiles—Database and tools, Nucleic Acids Res. 33 (2005) D562–D566 (Datafsbase issue).
[76] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G.G. Lara, E. Holloway, M. Kapushesky, et al., ArrayExpress—A public repository for microarray gene expression data at the EBI, Nucleic Acids Res. 33 (2005) D553–D555 (Database issue).
[77] C.A. Ball, I.A. Awad, J. Demeter, J. Gollub, J.M. Hebert, T. Hernandez-Boussard, H. Jin, J.C. Matese, M. Nitzberg, F. Wymore, et al., The Stanford Microarray Database accommodates additional microarray platforms and data formats, Nucleic Acids Res. 33 (2005) D580–D582 (Database issue).